DOCUMENT RESUME

ED 445 013 TM 031 550

AUTHOR Impara, James C.; Giraud, Gerald; Plake, Barbara S.
TITLE The Influence of Providing Target Group Descriptors When Setting a Passing Score.
PUB DATE 2000-04-00
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Definitions; *Groups; High Schools; Middle Schools; *Pass Fail Grading; *Secondary School Teachers; *Standards
IDENTIFIERS *Standard Setting

ABSTRACT

A study was conducted to explore empirically the effect of different definitions of the target examinee on the judgment of panelists setting a passing score. Two cut score studies were done in a school district for the same test within a 6-month period, and different definitions of the target candidate were provided for each study. In October 1998, 15 teachers, all of whom were middle school or high school mathematics teachers, participated in the first standard setting exercise. The target student was called the "Barely Master Student," a student whose skills are sufficient to justify graduation, "but just barely." The methods and procedures for the April study were essentially the same, but the target examinee was designed to be more skilled. The 10 panelists who participated in both studies typically set higher standards when the target student definition represented a more skilled student. The findings suggest that under the same conditions, cut scores can be very reliably set across time and panels, but when elements of the cut score study change, the cut score may change accordingly. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 013

The Influence of Providing Target Group Descriptors when Setting a Passing Score

James C. Impara, Ph.D.

Gerald Giraud, Ph.D.

Barbara S. Plake, Ph.D.

University of Nebraska-Lincoln

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Mills, Melican, and Ahluwalia (1991) advocate strongly for having panelists who participate in standard setting studies reach a consensus about the characteristics of the examinees who are at the cut score (i.e., the definition of minimal competence). In reaching consensus, they advocate the development of a set of behavioral characteristics that can be used to describe these target candidates. Mills, et al. suggest that agreement among panelists will result in a more defensible cut score, especially when using a test-centered model (e.g., Angoff, 1971) for setting the cut score. Their rationale is that if all panelists have similar examinees in mind, then panelists will have a reference set of behavioral characteristics to use when examining and judging individual test items.

Berk (1996) reinforces this need for extensive training. He cites others who have made similar statements (e.g., Melican and Mills, 1986; Mills, Melican, and Ahluwalia, 1991). Reid (1991) although offering no evidence that training is effective, suggests that effective training will lead to item judgments that are stable over time. Berk (1996) offers a set of criteria to be used in evaluating the

TM031550

BEST COPY AVAILABLE

validity of the standard setting process. One criterion he proposes is the inter- and intra-judge consistency of standards. A similar criterion is proposed by Kane (1994) who suggests that this reliability concern is a necessary condition for providing evidence of the validity of the cut score process.

Several studies have looked at both the impact of training judges in the description of the target candidate, and in the extent inter and intra judge consistency can be attained. For example, Fehrmann, Woehr, and Arthur (1991) examined the consistency and reliability of the cut score under different conditions of defining the target examinee. They found that panelists who were either provided with a well-defined frame of reference, or who developed a consensus-based definition (both strategies are consistent with the methods proposed by Mills, et al., 1991) produced a more consistent and reliable cut score than did a group who had no frame of reference.

More recently Plake, Impara, and Irwin (1999) reported on a study in which panels of judges set cut scores on tests consisting of a small subset of the same items across two years. Using only the set of common items the two panels from one year to the next set essentially the same cut score. Some panelists were on both panels and these panelists were split out from the total group and the repeating panelists and first time panelists were analyzed separately. Both groups set essentially the same cut score on these items in the second year and the panelists who had been on the previous year's panel set essentially the same cut score on these items both years. This suggests that given the same training and the same definitions of the target candidate, that panelists can demonstrate

both inter and intra judge consistency. The methods recommended by Mills, et al. (1991) to define the target examinee were followed in both years.

Giraud (1999) examined qualitatively how panels composed of public school teachers attended to the process and training in several standard-setting studies that employed a modified Angoff method. The teachers indicated that the process of defining the target examinee was very important and they referred back to the behavioral definitions frequently when making item-rating judgments. Teachers indicated they were able to set aside their personal judgments of what constituted competence in order to comply with the specific definition provided by the school system (as presented by the study facilitators).

These studies suggest a) that the judges attend to the training and were sensitive to the definition of the target examinee, and b) that similar training and definitions will result in judges making equivalent judgments on separate occasions. Of interest is whether providing different definitions of the target examinee in the context of training will result in judges making different judgments concerning the performance of the target examinee.

The purpose of the present study was to examine empirically the effect of different definitions of the target examinee on the judgment of panelists. This study was done in a medium-sized mid western school district. Two cut score studies were conducted for the same test within a six-month period. Different definitions of the target candidate were provided for each study. Approximately two-thirds of the panelists in the first study also participated in the second. The procedures and definitions of the target examinees are described below. The results of the two cut score studies are described and discussed

Methods and Procedures - October, 1998

In October, 1998 a panel of 15 teachers participated in a standard setting study in which a modified Angoff (1971) method was used to set a cut score. These 15 teachers (all of whom were middle or high school mathematics teachers were selected (and who volunteered after being selected) by the school system such that they collectively represent a cross section of the district's teachers and their classes represent a cross section of the district's students. An attempt was made to insure that all participating teachers had at least three years of teaching experience and at least two years (the current year and last year) teaching mathematics to ninth grade students.

The *High School Mathematics Proficiency Examination* used as a graduation examination in mathematics is in its third version. The current version of the examination is believed to be of sufficient psychometric quality to be used as one way for students to demonstrate their readiness, in mathematics, to graduate from high school. The test development efforts over the past several years have involved many individuals (teachers, specialists, and parents) in the school system. The objective of administering this test is to classify students into only two categories: those who demonstrate mathematics performance sufficient to graduate from high school and those whose performance is insufficient.

The content of the mathematics test includes 15 multiple-choice items, five short-answer items (all are computational and require students to show their work), and three long-answer items (all computational, multi-step problems, that

require students to show their work). The test is administered over a two-day period; Parts 1 and 2 (the multiple-choice and short-answer questions) are administered in a 45-minute period the first day. Part 3 is administered in a 30-minute time frame on the second day. Items in each part are weighted such that each part carries equal weight in computing the total score. Specifically, the 15 multiple-choice items each count 2 points, the 5 short-answer items each count 6 points, and the three long-answer items each count 10 points.

The modified Angoff method entails using teachers to examine each item on the test and estimate how a typical "Barely Master" student will perform on that item. The training is described followed by a description of the process used to make performance estimates for multiple choice items. That description is followed by a description of the variation in the method used for constructed response items as it was used in this study.

Training of panelists

The panelists were provided an overview of the activities to be done, including an overview of the Angoff to be used in this study. Once the panelists indicated that they understood the basic structure of the standard setting method, the initial training for the panelists began. The initial training consisted of providing the teachers with a description of the target student that was adopted by district staff to be acceptable. This target student was called the Barely Master Student (BMS), and was defined as follows: "The student can complete some appropriate mathematical tasks independently and can get by on other tasks with normal help from the teacher or other adult. This student is one

who can do most assigned tasks, but only after careful introduction, help in some problem solving steps, and considerable effort on the student's part. The student's skills are sufficient to justify graduation, but just barely."

After presenting this definition to the teachers, they were asked to visualize a specific BMS with whom they have interacted and to describe what characteristics would make mathematics tasks easy or difficult for that student. All panelists indicated they had such a student in mind for this exercise. The purpose in seeking descriptions of the BMS in mathematics at grade 9 was so that all panelists would have a common understanding of such students. Easy and difficult tasks were listed for each topic within the Table of Specifications (TOS) independently. The majority of characteristics listed were descriptions of specific skills that might be tested. An example of an easy task was: Perform direct single step word problems. An example of a difficult task was: Perform multi-step word problems.

Following this development of a behavioral description of the target student, teachers were provided a set of practice items on which the Angoff procedure was practiced. These practice items, six multiple choice and 2 short answer, had been administered in the school system and had item performance characteristics (p-values) similar to the operational items.

Angoff Ratings for Multiple Choice Items

For the High School Mathematics Proficiency Examination multiple choice items, teachers were asked to conceptualize (after a training activity) a specific barely master student they had taught. Keeping this student in mind, they were

directed to indicate, for each multiple-choice item, whether the student they had in mind would answer the item correctly or not (Yes or No) as described in Impara and Plake (1997). This was done for each multiple-choice item the teachers rated. After an initial rating, actual performance data from a sample of almost 800 students was provided to the teachers and they were asked to make a second estimate that could be either the same or different from their first estimate (the data provides a reality check to ensure that expected performance is not set either unrealistically high or low because the teacher has misjudged how hard or easy the item actually is). The cut score is based on the second estimate. It is calculated by summing, for each teacher, the number of "yes" items and then averaging these values across the teachers.

Angoff Ratings for Constructed Response Items

For the constructed response items, a paper selection method was used. This method required panelists to read a set of anchor papers (described below) and identify the papers that either represent the performance of the target student (the barely master) or, if no papers are shown that exactly match that performance, then two papers are selected that "bracket" the performance of the target student. Teachers were not advised of the scores on the anchor papers. The teachers were then provided actual performance data regarding the average score for students on that question, or set of questions, and also the cumulative percent of students at each of the score points (an estimate of the percent of students who would pass or fail at each score point). After being provided data on the average score and the distribution of scores, teachers were asked to review

the anchor papers a second time and make a final decision about which papers either represented, or bracketed, the performance of the target student. The cut score for an item was the average score across all teachers for the papers they selected in their second ratings.

The district staff selected anchor papers for each constructed response item. Specifically, a set of anchor papers was selected for each item prior to the standard setting workshop. At least two papers at each score point were provided. The anchor papers were selected such that they met the following criteria:

1. Selected papers are representative of student responses at that score point. That is, if most students made a particular error that resulted in a particular score, then that error should be reflected in the anchor papers. If a specific score resulted from several different answer strategies, then as many of the different answering strategies should be represented as possible. The occurrence of multiple strategies was most likely to occur in the mid-ranges of scores, as very low or very high scores were mostly all correct or all incorrect.
2. Selected papers are scored correctly and accurately. The basis for scoring was not to be an issue.
3. Selected papers are written legibly and darkly enough that they could be photocopied.

Methods and Procedures – April, 1999

The methods and procedures for the April study were essentially the same as those used in October. The differences were in the way the target student was described and in the number of panelists. Teachers were not told what their individual cut scores were in either October or April.

The school district staff wanted to change the definition of the BMS for the April meeting. They felt the definition used in October was targeted too low and that it would be difficult to justify the outcome of that definition in the context of using the results of the test as a high school graduation demonstration at grade 9. The school district central office staff provided a new definition of the BMS. The new definition of the BMS was: "The barely master student is able to solve some multi step application problems using a numerical or 'brute force' method, but has difficulty using traditional algebraic methods. The barely master student can solve most algebraic, geometric, or simple arithmetic applications that are not embedded in context (e.g., percent, proportion, probability, mean). The 'Typical' barely master student is in ninth grade algebra and demonstrates the skills necessary to earn a grade of 'C+' or 'C'. Or, the 'typical' barely mast student is a very strong transition math student, e.g., a student who demonstrates the skills necessary to earn a grade of 'A'." It is important to note that the district staff, who developed this definition, intended it to describe a more skilled examinee than the previous definition. This perception seems to have been shared by the teachers who, for example, described the April BMS as a student who is able "to

solve some multi step application problems" which was one of the behaviors an October BMS would find difficult.

In April there were 20 panelists. These panelists were selected using the same criteria that were used in October. Ten of the panelists in April had also participated in the October study. It is the data from these ten panelists that is of interest in this paper.

Results and Discussion

This study was intended to examine the extent that teachers would set a consistent cut score when provided different definitions of the target examinee, but when the purpose of the test is ostensibly the same. The perception of the school district staff was that that the second definition (for the April study) defined a more skilled student and would result in a higher cut score, resulting in more students scoring below the cut.

The expectation that both the new and the repeating panelists (i.e., the panelists who participated in both October and in April) would set a higher cut score in April was realized. Table 1 shows the cut scores for both October and April for all panelists combined, for the panelists who participated in only one study, and for the repeating panelists.

The October cut score for the 10 repeating panelists was 50.20 (SD=5.9). In April, just six months later, these panelists set a higher cut score 55.80 (SD=11.64). Also of interest was the almost 4-fold increase in the variance from October to April for all panelists, but which was greatest for the repeating panelists. The new panelists in April also set a higher cut score than did the non-repeating panelists in October.

Table 1 about here

The repeating panelists are typically setting higher standards when the target student definition is representing a more skilled student. Therefore either the higher standard set in April was due to the panelists being basically more stringent, without regard to the target candidate skill level or because the panelists were being sensitive to the higher skill level of the target candidate. Data from the repeating panelists suggests two possible explanations. First, we can infer that either the repeating panelists set different performance standards as a function of the different target skill level as defined in the training. Second, that the repeating panelists were familiar with test items and may believe the items are of relatively low difficulty without regard to changes in the definition of the target examinee. The second option can be discounted for two reasons. First, the work reported by Plake, et al. (1999) suggests that panelists who repeat and are provided consistent definitions of the target examinee set the same cut score. Therefore, it does not appear that repeat exposure to the same items distorts panelists' estimates of item difficulty for the target examinee. Second, the panelists in both October and April were provided with the same item p-values for the total group of examinees, thus the item difficulties were known for at least the average 9th grade student across the school district (including transitional, algebra, and a few geometry students).

As shown in Table 2 only three repeating panelists had lower values in April. One of these three was the panelist with the lowest value in October and this panelist had an even lower cut score in April (moving from a cut score of 38 to a cut score of 32. The largest change was exhibited by panelist 2, whose April

cut score was 22 points higher than his/her October cut score (moving from 44 to 66). Panelist 2 was accompanied by two other panelists who made a substantial change. We presume these changes were made as a result of the revised definition. It would have been interesting to interview these ten panelists after the study to learn more directly why they made such dramatic changes. These wide swings, as shown in Table 2, resulted in the substantially larger variance in April than in October.

Table 2 about here

Conclusions

Teachers who participate in a cut score study on the same test but at different times using different "mastery" definitions set a different cut score. Giraud (1999) found that teachers indicated they attended carefully to the definition of the target examinee. Moreover, he reported that discussion about the behavioral characteristics of that examinee were helpful to teachers in making decisions about their estimates of item performance. Intensive training has been advocated (Mills, et al., 1991, Fehrmann, et al. 1991) and such training clearly makes a difference in how teachers perceive of the target examinee. Because these teachers were not asked explicitly if they perceived that the April definition was more rigorous than the October definition of the target examinee, it is not clear how all the teachers interpreted the definition of the Barely Master Student on each occasion. Three of the teachers may have felt the October definition was the more rigorous because their October cut scores were higher than the cut scores they set in April. The remaining teachers, however,

appeared to interpret the new definition as more rigorous and they set a higher cut score.

We can conclude from this that under the same conditions (definitions, methods) cut scores can be very reliably set across time and panels (Plake, et al., 1999), when elements of the cut score study change, then the cut score may change accordingly. Thus, cut scores are context dependent and constructed based on judges' interpretation of the definition provided and by the discussion that fleshes out that discussion in more behavioral terms. There may also be a sensitivity of the judges to the desires of the district regarding the impact of the cut score. One interpretation of these findings is that perhaps cut scores are a product of both expert deliberations and the nature and content of the training these experts receive. Facilitators of cut score studies need to be sensitive to the definitions used in the training and to the processes used to obtain behavioral descriptions of the target examinees.

References

Angoff. W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Edition, 508 – 600. Washington, DC: American Council on Education.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). Applied Measurement in Education, 9 (3), 215-235.

Fehrmann, M. L., Woehr, D. J., & Arthur, W. (1991). The Angoff cutoff score method: ; the impact of frame-of-reference training. Educational and Psychological Measurement, 51, 857-872.

Giraud, G. (1999). Making the cut: A qualitative inquiry of teachers in the standard setting process. Unpublished doctoral dissertation, University of Nebraska – Lincoln.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. Journal of Educational Measurement, 34, 355-368.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64 (3), 425-461.

Melican, G. J., & Mills, C. N. (1986, April). The effect of knowledge of item difficult for selected items on subsequent ratings of other items using the Angoff method. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. Educational Measurement: Issues and Practice, 10 (2), 7 – 10.

Plake, B. S., Impara, J.C., & Irwin, P. (1999). Validation of Angoff-based predictions of item performance, ERIC Clearinghouse on Assessment and Evaluation, TM029717

Table 1. Cut scores from the total panel and the non-repeating and repeating panelists for both October and April.

Panel	Cut Score (St. Dev.)	Median	Min/Max
Oct. – Full (N = 15)	49.30 (6.17)	51.50	38/56
Oct. – Non repeaters	47.75 (7.41)	49.50	38/54
Oct. – Repeaters only	50.20 (5.90)	51.50	38/56
Apr. – Full (N = 20)	52.80 (10.84)	53.00	31/72
Apr. – Non repeaters	49.80 (9.62)	52.00	31/61
Apr. – Repeaters only	55.80 (11.64)	57.50	32/72

Table 2. October and April cut scores for repeating panelists

Panelist #	October	April	Change
1	44	66	22
2	54	72	18
3	51	67	16
4	52	58	6
5	50	47	-3
6	56	57	1
7	38	32	-6
8	46	50	4
9	55	59	4
10	56	50	-6

Correlation between October and April = .51